A note on construction of heuristically optimal Pena's synthetic indicators by the particle swarm method of global optimization

SK Mishra

North-Eastern Hill University, Shillong (India)

24. March 2012

A Note on Construction of Heuristically Optimal Pena's Synthetic Indicators by the Particle Swarm Method of Global Optimization

SK Mishra
Department of Economics
North-Eastern Hill University, Shillong (India)
e-mail: mishrasknehu@hotmail.com

Abstract

Pena's method of construction of a synthetic indicator is very sensitive to the order in which the constituent variables (whose linear aggregation yields the synthetic indicator) are arranged. Due to this, Pena's method can at present give only an arbitrary synthetic indicator whose representativeness is indeterminate and uncertain, especially when the number of constituent variables is not very small. This paper uses discrete global optimization method based on the Particle Swarms to obtain a heuristically optimal order in which the constituent variables can be arranged so as to yield Pena's synthetic indicator that maximizes the minimal absolute (or squared) correlation with its constituent variables.

JEL Code: C18, C43, C44, C61

Keywords: Synthetic indicators, Pena's distance, Particle swarm, Discrete Global Optimization, Composite indices, Maxi-min absolute correlation

I. Introduction: A synthetic indicator (or composite index), Z(n), is an n-element array that represents a multitude of other n-element arrays (called constituent variables), X(n, m), such that Z is a mapping of X. In this description, there are two point to note: first that Z represents X or a significant part of information content of X is preserved in Z and the second that there is a rule that establishes a correspondence between Z and X. Very often, Z is a linear combination of X (such that Z=Xw or aggregation of weighted X). Also, on many occasions, the degree of representation is measured by the coefficient of correlation, $r(Z, x_j)$, between Z and $x_{.j} \in X$.

There are, indeed, many and varied methods to construct Z from X (Munda & Nardo, 2005; Mishra, 2007; Mishra, 2009; Mishra, 2010b; Pena, 1977; Somarriba & Pena, 2009). The determination of weights (w) could be subjective, extraneous, intrinsic, etc. While subjectively chosen weights are based on opinion, impression, etc, the objective weights could be based on extraneous variables, Y (while Y $\not\subset$ X), and intrinsic weights are derived from X itself. The Human Development Index (HDI), for example, is an index that uses subjective weights (supported by the logic of insufficient reason to assign different weights to different variables); it is composed of three variables, life expectancy (LE), Educational achievement (ED) and per capita income (PCI) each assuming equal (1/3) weight. The Walsh price index, for example, is the weighted sum of the current period prices divided by the weighted sum of the base period prices with the geometric average of both period quantities serving as the weighting mechanism (or $P_{iW} = [\sum P_{it}.(q_{i0}.q_{it})^{1/2}] / [\sum P_{i0}.(q_{i0}.q_{it})^{1/2}]$. Here $(q_{i0}.q_{it})^{1/2}$ is used as the weight for the price of i^{th}

commodity. Among the methods that derive weights intrinsically (from X itself), the Principal Component Analysis (PCA) is perhaps most popular. PCA obtains Z=Xw such that $\sum_{j=1}^{m}|r(Z,x_j)|^2$ is maximized. This amounts to maximizing the Euclidean norm $\left[\sum_{j=1}^{m}|r(Z,x_j)|^2\right]^{0.5}$. Analogously, one may derive weights by maximization of the absolute norm, $\sum_{j=1}^{m}|r(Z,x_j)|$ or the Chebyshev norm ($L_{p\to(-\infty)}$), $\min_j(|r(Z,x_j)|)$. Obtaining weights and the synthetic indicators in these manners would not be called PCA, but they run parallel to PCA and are often more inclusive than the PCA, which is highly elitist (Mishra, 2007, 2011).

II. Pena's Distance and Method of Constructing Synthetic Indicators: Pena (1977) proposed a new method of construction of synthetic indicators based on his concept of distance (DP2) defined as:

$$DP2_i=\sum_{j=1}^{m}\left[\left(\frac{d_{ij}}{\sigma_j}\right)\left(1-R^2_{j,j-1,\ldots,1}\right)\right];i=1,2,\ldots,n \qquad \ldots \qquad (1)$$

where: $i=1,2,\ldots,n$ are cases (e.g. countries); m is the number of constituent variables, X , such that $x_{ij}\in X;i=1,2,\ldots,n;j=1,2,\ldots,m$; $d_{ij}=\left|x_{ij}-x_{rj}\right|;i=1,2,..,n;j=1,2,\ldots,m$; r is the reference case; σ_j is the standard deviation of constituent variable j ; $R^2_{j,j-1,\ldots,1}$; $j>1$ is the coefficient of determination in the regression of x_j over $x_{j-1,}x_{j-2},\ldots,x_1$. Moreover, $R^2_1=0$ (Somarriba & Pena, 2009). A synthetic indicator constructed by Pena's method is claimed to have almost all desirable properties (Pena, 1977; Zarazosa, 1996; Somarriba & Pena, 2009; Montero et al., 2010; Garcia et al., 2010; Martína & Fernández, 2011).

However, it has been demonstrated (Mishra, 2012) that an application of Pena's method of construction of synthetic indicators suffers from indeterminacy and arbitrariness. This is because of the fact that the weight ($w_j=1-R^2_{j,j-1,\ldots,1}$) obtained by the j^{th} (standardized) constituent variable, d_{ij}/σ_j , depends on its position in the order or the value of j . Thus, if there are 10 constituent variables, they can be arranged in 10-factorial ways and we will have 3628.8 thousand possible synthetic indicators (differing from each other). For 25 constituent variables we may construct about 1.55112E25 synthetic indicators. From such a large number of indicators, it is impossible to choose the one that represents the constituent variables best. As a result, Pena's method as applied today is arbitrary and considering such a synthetic indicator better than those constructed by other methods is a matter of unfounded belief (Mishra, 2012).

III. The Objective of this Paper: Choosing the best representative Pena's synthetic indicator while the number of constituent variables is not very small hinges on computing the synthetic indicators for every permutation or the order in which the variables enter in the formula (eq. 1). Clearly, this is a practically impossible task if one goes by constructing the indicator for every permutation of constituent variables and choosing the best (yet undefined) from among them. Therefore, we must find out a method which can be applied to obtain the best (or near-best) synthetic indicator in practice. This can be achieved by optimization of Pena's indicator on some acceptable criterion. This paper is an attempt in the same direction.

IV. The Criterion of choosing the Best Synthetic Indicator: As it has been pointed out earlier, the PCA criterion of 'best' is maximization of the Euclidean norm of the coefficients of correlation between the synthetic indicator and the constituent variables (or, in practice, maximization of $\sum_{j=1}^{m}|r(Z,x_j)|^2$). As a consequence, PCA-based synthetic indicators ignore (or assign marginal weights to) those constituent variables that correlate poorly to the leading (elite) variables. On the contrary, the choice of the Chebyshev norm (or Minkowsky's $L_{p\to(-\infty)}$ norm) yields maximin solution, that maximizes the minimum (absolute) correlation, $\min(|r(Z,x_j)|)$. . This criterion yields most inclusive synthetic indicator that assigns suitable weight to weakly correlated variables also. Therefore, we propose in favour of choosing the criterion as maximization of the Chebyshev $L_{p\to(-\infty)}$ norm.

V. The Method of Optimization: Maximization of $\min(|r(Z,x_j)|)$ is not amenable to the traditional methods of optimization, especially in view of that fact that in Z=Xw, the weights, ($w_j = 1 - R^2_{j,j-1,\ldots,1}$), depend on the order in which the constituent variables enter into the formula. This poses a combinatorial problem. It has been found that the methods of global optimization, such as the genetic algorithms (Holland, 1975; Wikipedia: Genetic Algorithm), the discrete particle swarm, the taboo search (Glover, 1989, 1990) and the ant colony algorithm (Dorigo, 1992) are appropriate and effective.

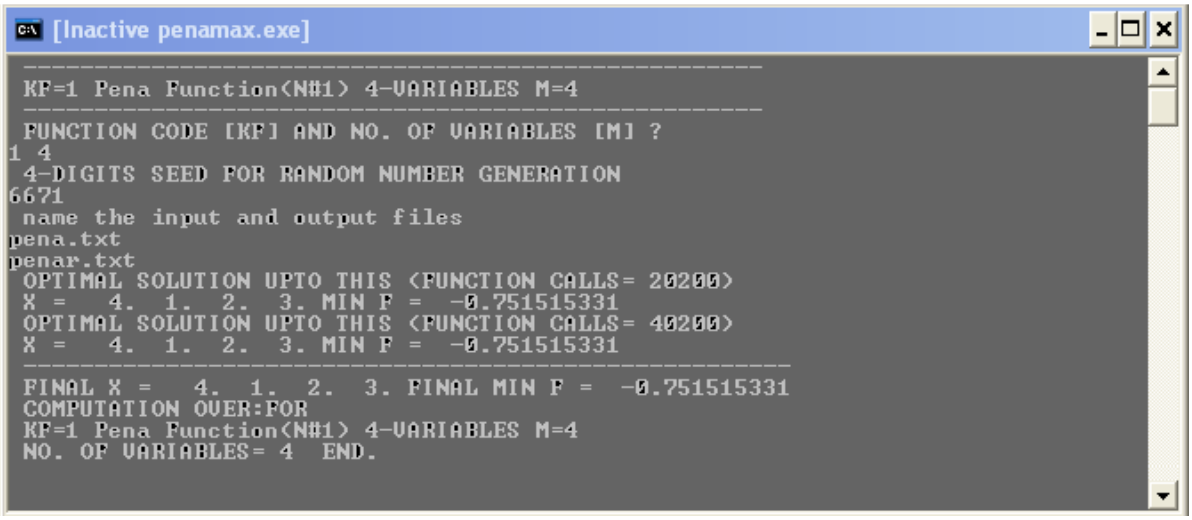
In this paper, we have chosen the discrete particle swarm method of global optimization for meeting the objective. The details of the particle swarm method (Kennedy & Eberhart, 1995) in the continuous parameter space are available in Bank et al. (2008) and Mishra (2010a). Discrete problems can well be optimized in continuous space through a suitable mapping of the problem space to the potential solutions generated by the particle swarm method. Parsopoulos & Vrahatis (2006) applied the Smallest Position Value (SPV) mapping mechanism (Tasgetiren et al., 2004) for solving the discrete optimization problem. In the SPV scheme the schedule is produced by placing the index of the lowest valued particle component as the first item, the next lowest as the second and so on in that order. For example, a given particle having the coordinate position (5.16, 3.15, 1.28, 2.17) would represent the potential schedule (3, 4, 2, 1). This potential schedule would then be submitted to the objective function for an assessment of its fitness (Bank et al., 2008). We have used the SPV method in this paper. This scheme will normally generate the schedule that will represent a non-degenerate coded permutation, since it is very unlikely that any two random numbers generated in the continuous parameter space will be equal. However, any two (or more) potential schedules generated by this method may be identical. To avoid this, embedding of the taboo search method in the particle swarm optimization algorithm is warranted. However, such an attempt has not been made presently and it is left to be pursued in the future research.

VI. The Test Data: Using the Human Development Report of UNDP, 2004 data and the additional information on the measures of inequality, Sarker et al. (2007), argued that Human Development Index (HDI) should include income equality measures (EQ) also in addition to the three conventional measures, viz. life expectancy (LE), education (ED) and per capita gross domestic product at the purchasing power parity with the US $ (PCI). The data are reproduced in Mishra (2012). We use these data/variables (LE, ED, PCI and EQ) to construct Pena's synthetic indicators.

By a complete enumeration of all 24 permutations (of 4 constituent variables) it has been found that the permutation (EQ, LE, ED, PCI) indexed as (4, 1, 2, 3) has the minimal absolute correlation, $\min(|r(Z,x_j)|)$

= 0.7515, which is maximal for all possible 24 permutations (Mishra, 2012). A successful search by the discrete particle swarm method should obtain this.

VII. The Findings: The discrete particle swarm successfully finds the permutation (4, 1, 2, 3) and the max($\min(|r(Z,x_j)|)$) = 0.7515. It tallies perfectly with the value obtained through complete enumeration.



VIII. Experiments with Some Artificial Data (X of Larger Dimensions): Combinatorial optimization in discrete parameter space is extremely time-consuming. To gauge into the prospects of using the discrete particle swarm optimization method for the problem at hand (identifying the heuristically best Pena's synthetic indicators), we have tested the method for X of 12 and 25 dimensions, i.e. X(125,12) and X(125, 25) and noted the time needed to obtain the solutions. The specification of the computer is Intel Core 2 Duo E4600 @ 2.4 GHz, which, by to-day's standard, is a slow machine. The program (available on request to the author) is written in FORTRAN-77 (compiled by Force3.0 compiler). Data (X) used for this purpose are arbitrarily generated. Our objective is not interpretation, but estimation of time required for finding the optimal (or near-optimal) solution. The results are presented in Table-1.

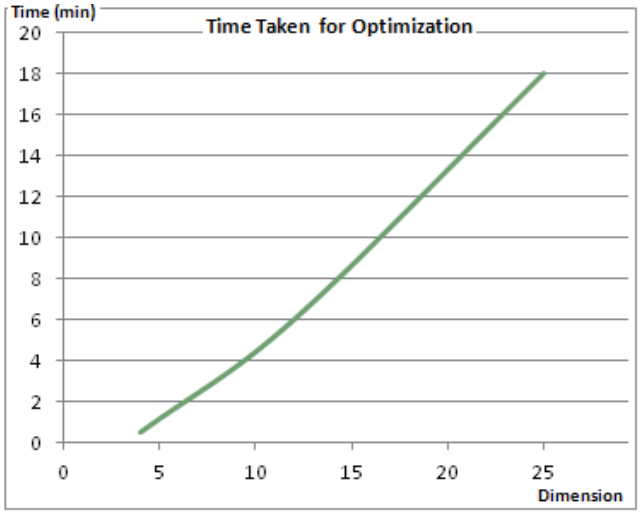
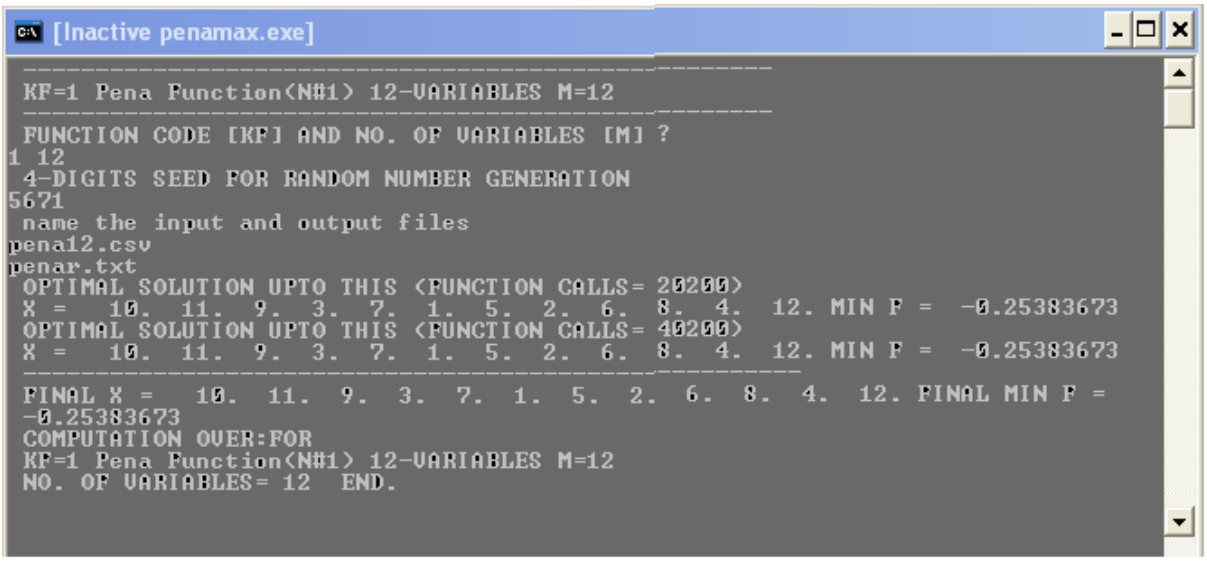
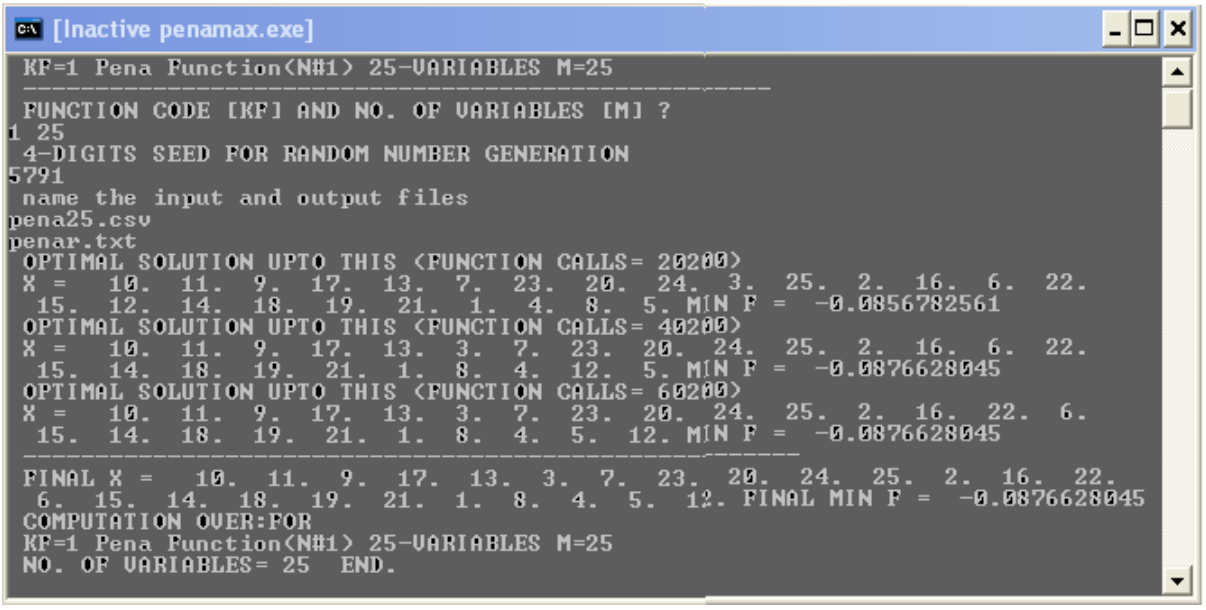


Table-1. Dimension of the Problem and Time taken for Solution [Machine: Intel Core 2 Duo E4600 @ 2.4 GHz]				
Sl. No of run	No. of cases (n)	No. of Variables (m)	Time Taken	Remarks
1	125	4	0.5 minute	1 (benchmark)
2	125	12	6 minutes	12 times
3	125	25	18 minutes	36 times





IX. Concluding Remarks: By way of conclusion we note the following:

- The otherwise unmanageable problem of finding the best (or near-best) synthetic indicators by Pena's method is made manageable by the discrete particle swarm method of global optimization. The time required for solution does not increase exponentially with the size of the problem.

- The taboo search method may be embedded into the particle swarm method (or it may be used directly) to make the optimization procedure more time-efficient. However, it requires experimentations with the taboo search method.
- The results obtained by the particle swarm method (or other methods of global optimization) may not be optimal, but only near-optimal, since these methods have a tendency to be caught into the local optimum trap. Several runs or fine-tuning of the optimization parameters may be required.
- Faster computing machines may greatly reduce the time required for computation.
- The global optimization methods are almost always amenable to parallel or multi-thread computing. This facility may be of a great relevance for solving the large-scale problems.
- The method is also amenable to changes in the norm used as a criterion of optimization.

References

1. Banks, A., Vincent, J., & Anyakoha, C. (2008): "A Review of Particle Swarm Optimization. Part II: Hybridisation, Combinatorial, Multicriteria and Constrained Optimization, and Indicative Applications", *Natural Computing,* 7(1): 109–124.
2. Dorigo, M. (1992) *Optimization, Learning and Natural Algorithms*, PhD thesis, Politecnico di Milano, Italie, 1992.
3. García, E.C., Rodríguez Martín, J.A. & Pabsdorf, M.N. (2010): "The Features of Development in the Pacific Countries of the African, Caribbean and Pacific Group", Social Indicators Research 99(3): 469-485.
4. Glover, F. (1989). "Tabu Search - Part 1". *ORSA Journal on Computing,* 1(2): 190–206.
5. Glover, F. (1990). "Tabu Search - Part 2". *ORSA Journal on Computing*, 2(1): 4–32.
6. Holland, J.H. (1975) *Adaptation in Natural and Artificial Systems*, Ann Arbor: The U. of Michigan Press.
7. Kennedy, J. & Eberhart, R. (1995). "Particle Swarm Optimization". Proceedings of IEEE International Conference on Neural Networks - IV : 1942–1948. doi:10.1109/ICNN.1995.488968.
8. Martína, J.A.R. & Fernández, J.A.S. (2011): "An Index of Maternal and Child Health in the Least Developed Countries of Asia", *Gac Sanit*. 2011 [in press]; doi:10.1016/j.gaceta.2011.05.021.
9. Mishra, S.K. (2007): "A Note on Human Development Indices with Income Equalities", http://ssrn.com/abstract=992854 or http://dx.doi.org/10.2139/ssrn.992854.
10. Mishra, S.K. (2010a): "Performance of Differential Evolution and Particle Swarm Methods on Some Relatively Harder Multi-modal Benchmark Functions", *The IUP Journal of Computational Mathematics*, III(1): 7-18.
11. Mishra, S.K. (2010b): "Construction of an Index: A New Method", in *Growth and Human Development in North-East India*, Nayak, P. (ed.), Oxford University Press: 24-35.
12. Mishra, S.K. (2011): "A Comparative Study of Various Inclusive Indices and the Index Constructed by the Principal Component Analysis", *IUP Journal of Computational Mathematics*, 4(2): 7-26.
13. Mishra, S.K. (2012): "A Note on the Indeterminacy and Arbitrariness of Pena's Method of Construction of Synthetic Indicators", *(unpublished paper) http://mpra.ub.uni-muenchen.de/37554/*

14. Montero, J.M., Chasco, C. & Lanaz, B. (2010): “Building an environmental quality index for a big city: a spatial interpolation approach combined with a distance indicator”, *J. Geogr. Syst.* 12: 435-459.
15. Munda, G. & Nardo, M (2005): “Constructing Consistent Composite Indicators: The Issue of Weights”, EUR 21834 EN, Institute for the Protection and Security of the citizen, European Commission, Luxembourg.
16. Pena, J. B. (1977): Problemas de la medicio´n del bienestar y conceptos afines. Una aplicacio´n al Caso Espan˜ol. (I.N.E.: Madrid).
17. Sarker, S., Biswas, B. & Soundrs, P.J. (2006): “Distribution-Augmented Human Development Index: A Principal Component Analysis”, GSP, College of Business, Utah State Univ., USA. www.usu.edu/cob/econ/graduatestudents/documents/papers/developmentpaper.pdf. (visited on April1 24, 2007).
18. Somarriba, N. & Pena, B. (2009): “Synthetic Indicators of Quality of Life in Europe”, *Soc. Indic. Res*. 94(1): 115–133.
19. Zarzosa, P. (1996): Aproximacio´n a la medicio´n del bienestar social. Valladolid: Secretario de Publicaciones.
20. Parsopoulos, K.E. & Vrahatis, M.N. (2006) “Studying the Performance of Unified Particle Swarm Optimization on the Single Machine Total Weighted Tardiness Problem” In: Sattar, A., Kang , B.H. (eds) *AI 2006, LNAI 4304*, Springer-Verlag: 1027–1031.
21. Tasgetiren, F., Sevkli, M., Lian, Y.C., & Gencyilmaz, G. (2004) “Particle Swarm Optimization Algorithm for Single Machine Weighted Tardiness Problem”, *Proceedings of IEEE Congress on Evolutionary Computation*: 1412–1419.